A comprehensive survey of the approaches for pathway analysis using multi-omics data

Supplementary Note

Zeynab Maghsoudi[1], Ha Nguyen[1], Alireza Tavakkoli[1] and Tin Nguyen[1,*]

[1] Department of Computer Science and Engineering, University of Nevada, Reno
∗Corresponding: tinn@unr.edu

Contents

1 CancerMA

CancerMA [1] is a web-based tool that provides a framework for meta-analysis of multiple cancer gene expression datasets, particularly microarray. By the time of writing this manuscript, CancerMA is not accessible.

In principle, CancerMA retrieves and processes (quality control, background correction, normalization) microarray data from ArrayExpress [2] and GEO [3] databases. Accordingly, this website has established its own microarray database, which contains 80 individual curated cancer expression matrices originating from 45 experiments and covers 13 different cancer types. For each matrix, the rows represent the gene/probe IDs, while the columns are samples labeling control or disease. Another local database is built for automatically mapping probe to gene, which contains gene and probe annotations curated from the Ensembl [4] database, HUGO Gene Nomenclature Committee (HGNC) [5] database, and the Affymetrix database. The website also gets pathway information from Gene Ontology (GO) [6] for pathway analysis.

The users need to upload a user-supplied gene list to start an analysis. This list is then used to filter the features from 80 individual matrices. Differential Expression analysis using *limma* [7] R package is performed in each filtered data matrix. For each gene, lists of p-value and log2 fold change (log2FC), one per data matrix, are returned. These p-values are also corrected using FDR.

For single array analysis, genes adjusted p-value < 0.05 and log2FC > 1 are selected as DE genes. For the meta-analysis, the results from 80 individual analyses are combined for each cancer type as well as for all cancer types. Stouffer's [8] method and weighted linear combination are applied to calculate meta-p-values and meta-log2FC, respectively. If multiple probes are mapping the same gene identifier, the probe with the most extreme log2FC value with its corresponding p-value is selected. The following thresholds select DE genes: absolute meta-log2FC > 1 or a confidence interval above zero, and a meta-p-value < 0.05.

After applying the meta-analysis at the gene level, CancerMA obtains a list of DE genes. CancerMA then uses ORA to calculate the p-value for each GO term. The output of CancerMA includes the list of GO terms and their p-values for both meta-analysis and the analysis of each dataset. CancerMA also visualizes the analysis results using Circos, forest, and Krona plots.

2 INMEX

INMEX (INtegrative Meta-analysis of EXpression data) [9] is a web-based platform that performs meta-analysis of multiple gene expression datasets. It also supports the integration of transcriptomics and metabolomics data. Each input file is formatted as a data table, in which rows represent genes/probes and columns represent samples. After uploading the data, the website allows users to visualize the data using box plot and principal component analysis (PCA) to detect outliers or to investigate whether the data need to be further normalized. INMEX supports four types of gene IDs: Entrez, RefSeq, GenBank, and Ensembl. In addition, 45 probeset IDs corresponding to 45 microarray platforms for humans, mice, and rats, are supported in this tool. Note that all metabolite IDs will be mapped to KEGG compound IDs.

In the first step of analysis workflow, INMEX uses the moderated t-test in the *limma* package to perform differential analysis and to identify the differentially expressed genes within each input dataset. Next, users can choose one of the three following modules to continue their analysis. First, the 'Pattern extractor' module visualizes a set of user selected expression profiles as heatmaps. Second, in the 'enrichment network' module, pathway enrichment analysis is performed by applying ORA or GSEA on each subset of input data matrices. Results are shown as a table of pathways with their p-values and an interactive network of pathways. On the generated network, users can interactively explore each node to see differentially expressed genes and their statistics.

Third, the 'GO analysis' module, using hypergeometric test, enables users to identify the enriched GO terms. The input for this module is a list containing genes and their corresponding statistics, which are combined from individual differential analyses. Briefly, using the results from the *limma* t-test, INMEX implements five different strategies to combine the results, including combining p-values (Fisher's and Stouffer's methods), combining standardized differences (effect sizes), combining rank orders, combining votes, and direct data merging. Combining rank orders is based on computing the fold-change for each gene within each dataset. Then, the computed fold-changes for each gene is multiplied to obtain a single rank for the gene. Combining vote strategy is based on identifying DE genes within each dataset. Next, the number of datasets in which a gene is detected as DE would be considered as a gene's vote. The direct data merging approach combines all datasets into one single dataset and uses this dataset

for further analysis. Combining p-values and combining standardized differences approaches are vulnerable against outliers. On the other side, combining rank orders is more robust in facing with outliers and larger variations in different studies. The vote-counting method is essentially platform-independent, as only the final DE gene lists are used, and as the last one, direct data merge should only be used when gene sets are similar. The results from the meta-analysis are presented in detailed tables containing statistics from individual differential analyses, as well as statistics using the selected combiners.

3 Active Pathways

Active Pathways [10] is an R package on CRAN that allows users to integrate multiple omics datasets for pathway analysis. The input includes two files. The first file is a matrix of p-values in which rows represent genes and columns represent different datasets or data types. The second file is the pathway annotation in the format of GMT file.

For each gene (row), the method combines the p-values across the datasets/data types (columns) using Brown's method (combining dependent p-values), which is an extension of Fisher's method (combining independent p-values). After this step, we receive a list of combined p-values – one per gene. Next, the method remove genes with high combined p-values (a default threshold of 0.1 is used). The remaining genes are considered DE genes. Next, a ranked hypergeometric test (ranked ORA) is performed to compute the p-values of the pathways. The method also adjust the p-values using the Bonferroni method. Pathways with adjusted p-values smaller than 5% are considered significant.

Independently, for each dataset/data type, the method also applies the same threshold of 0.1 to determine the DE genes and then perform pathway analysis using ranked ORA to calculate the p-values of the pathways. The same adjustment method (Bonferroni) and threshold are applied to identify the pathways that are significant in each dataset/data type. The output of the method is a list of pathways that are significant in at least one of the analyses. The method reports a table that indicate which pathway is significant in which analysis (all individual analyses and the integrative analysis). This table serves as the input of another module, named Enrichment Map [11] from Cytoscape software package [12], for the visualization of pathway analysis and data integration.

4 MarVis-Pathway

MarVis-Pathway [13] is a standalone software implemented in MATLAB. The tool is designed with an emphasis on integration analysis of non-targeted metabolomics data and other omics data, including DNA microarray-based transcriptomics. The input is multiple .csv files in which rows are genes and columns are samples.

After importing the metabolomics and transcriptomics datasets in MarVis-Filter, each dataset's features are filtered. The filtration process starts with performing a signal-to-noise ratio test. Then the sample labels are permuted 1000 times (assignments of samples to conditions), and finally, the features with a false discovery rate (FDR Benjamini and Hochberg [14]) threshold of 0.05 is filtered. Each filtered gene set is stored in the MarVis-Filter clipboard. Finally, all filtered metabolomics and transcriptomics sets are combined by concatenating the corresponding data tables. The features of the combined set are mapped to metabolite and gene entries in the corresponding pathways from KEGG and BioCyc [15] database collection. The mapped features from the metabolomics datasets to metabolite entries are based on the corrected accurate masses. The transcriptomics features are mapped to gene entries using the corresponding IDs. The combined dataset is utilized by MarVis-Cluster module for visualizing the clustering based on 30 prototypes. The proportion of each type of data (including transcriptomics and metabolomics) is shown beside each cluster.

The pathway analyzer in this tool can apply three types of enrichment analysis on each gene set, including Entry-based, marker/feature-based, and sample-based analysis. The Entry-based is analogous to the ORA method, in which the number of entries in a pathway matched by the selected features (in MarVis-Filter or MarVis-Cluster) in comparison to the number of entries that could be matched over all pathways, is evaluated based on a hypergeometric distribution. In the marker/feature-based analysis, the calculated features' ranks in the MarVis-Filter module are considered for finding match entries in a particular pathway using a static or iterative hypergeometric test [16], a rank-sum, or a Kolmogorov-Smirnov test [17]. The third case, sample-based analysis, operates similarly to GSEA. The ranks of features are utilized beside a rank-sum or Kolmogorov–Smirnov test statistic. After calculating the

corresponding p-values for each input set for a specific pathway, the p-values will be combined using Fisher's [18] or Stouffer's method.

5 BLMA

BLMA (Bi-level meta-analysis) [19] is an R package available on Bioconductor. It allows users to perform meta-analysis of multiple mRNA datasets. The input is multiple matrices/datasets with grouping information (diseases vs. controls). For each dataset, BLMA allows users to perform pathway analysis using one of the following methods: over-representation analysis (ORA), Gene Set Analysis (GSA) [20], Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) [21], Impact Analysis (IA) [22]. After analyzing each dataset, each pathway has multiple p-values – one per dataset. For each pathway, users can combine the p-values using one of the following methods: addCLT (additive method), Fisher's, Stouffer's, minP, and maxP methods. The output is a list of pathways with the p-values obtained from each dataset or from the integration. The tool also adjusts the p-values using FDR. Note that addCLT is a method developed based on the Central Limit Theorem and the additive method. When the number of datasets are smaller than 20, the additive method is performed for calculating the combined P-values. It is proven that by increasing the number of studies, the additive method does not work properly. Therefore, the Central Limit Theorem (CLT) is suggested in the case that the number of studies is greater than 20.

6 DANUBE

DANUBE (Data-driven meta-ANalysis using UnBiased Empirical distributions) [23] does not provide an R package. This method is a meta-analysis approach that is able to correct for method bias and integration of multiple mRNA datasets. The method is based on the fact that pathway analysis often provide results that are biased toward well-studied conditions (cancer, Alzheimer's disease, etc.). The authors start by providing proof that the p-values obtained under the null hypothesis are not uniformly distributed. To do this, they downloaded a large number of control samples and then construct the datasets under the null by randomly label the samples as controls and diseases. They proved that the p-values are not uniformly distributed, and that the bias is due to the methods rather than the data.

The input of the method is multiple mRNA datasets with sample groupings. DANUBE uses one of the four methods, GSEA, GSA, IA, and PADOG, to perform pathway analysis in each dataset. It then uses the empirical distributions to correct the p-values obtained for each pathways. After bias correction, each pathway has multiple p-values – one per datasets. For each pathway, the method combines the p-values using the addCLT method, which is a combination of the additive and the Central Limit Theorem. The method also adjusts the p-values using FDR. The output if a list of pathways and their p-values (from individual dataset and from the aggregated analysis).

7 Mergeomics

Mergeomics [24] is available as a Bioconductor package and a web-based application. The method consists of two major modules. The first module, Marker Set Enrichment Analysis (MSEA), focuses on identifying significantly enriched pathways, whereas the second module, weighted Key Driver Analysis (wKDA), identifies genes that are the potential key drivers of the enriched pathways. Although these modules are introduced as sequential steps in Mergeomics, they can be performed independently as per user's interest.

The input of MSEA includes: (1) a matrix with two columns, marker IDs (e.g., genes) and associated minus log p-values or log fold change, (2) a matrix measuring the correlation between the marker IDs, (3) a list of functionally defined gene sets (e.g., biological pathways or co-regulated genes), (4) mapping between marker IDs and genes. The null hypothesis is that the differentially expressed (DE) genes are sampled from the same general population as genes from the reference set. Therefore, if there are more DE genes in a pathway than what can be expected by chance, then the pathway is likely to be enriched. To test this assumption, a chi-square-like statistic is calculated, and a permutation test (by randomly shuffling the gene or marker labels) is performed to retrieve the p-value for each pathway. The output from MSEA is a ranked list of pathways and their p-values. MSEA also allows users to perform a meta-analysis of multiple datasets (meta-MSEA). meta-MSEA first calculates the p-values obtained

from multiple datasets for each pathway and then combines these p-values using Stouffer's method. The combined p-values are then adjusted using FDR.

The input of wKDA is the set of significantly enriched pathways obtained from MSEA. By default, pathways with adjusted p-values less than 5% are selected as significantly enriched. If there is no significant pathway, the top 10 pathways will be used for wKDA. Users also need to upload a file that defines the topological structure of the pathways in the format of weighted graphs. The file has three columns: head node, tail node, and weight. The wKDA then uses this information to project enriched pathways onto a network model of gene regulation, thus returns the key driver gene as well as the number of all possible key drivers for each pathway. The enrichment p-values and their FDR adjusted values for these key drivers are also printed out in the form of a table.

8 mirIntegrator

The mirIntegrator tool [25] is available as an R package on Bioconductor. This tool allows users to perform the integration of miRNA and mRNA data for pathway analysis and disease subtyping. In order to perform topology-aware analysis, the software augments KEGG and Reactome pathways with miRNA-mRNA interactions available at miRTarBase [26] or TargetScan [27]. By default, the software provides a set of augmented pathways from the biologically validated interactions at KEGG, Reactome, miRTarBase, and TargetScan.

The input of mirIntegrator is a list of DE genes and their fold changes. The tool calculates the p-values of the pathways using ORA and Impact Analysis (IA). However, the generated augmented pathways by mirIntegrator can be utilized by many existing pathway analysis methods. In the paper, authors mentioned ORA and Impact Analysis (IA). IA is based on two types of evidence: i) the over-representation (ORA) of DE genes in a pathway, and ii) the perturbation factor of such a pathway, as measured by propagating expression changes through the pathway topology. The p-values for this evidence are calculated independently and then combined using Fisher's method. The output lists pathways sorted according to their nominal and adjusted p-values. This package also generates graphical representations of the augmented pathways, which can provide a more comprehensive view.

9 multiGSEA

The method multiGSEA [28] has an R package on Bioconductor. This method allows users to perform data integration and pathway analysis using pathways obtained from eight databases (implemented by *graphite* package [29]): PharmGKB [30], NCI/Nature Pathway Interaction Database [31], HumanCyc [32], SMPDB [33], Panther [34], Biocarta [35], KEGG [36–38], and Reactome [39]. The input of multiGSEA includes three data types: transcriptomics, proteomics, and metabolomics data. User-defined feature mappings map supported omics layers to the genes of the pathways. Transcriptomics and proteomics features can be mapped to the following formats: Entrez IDs, Uniprot IDs, Gene Symbols, RefSeq, or Ensembl IDs. Metabolomic features can be mapped to Comptox Dashboard specific IDs (DTXSID, DTXCID), CAS numbers, Pubchem IDs (CID, SID), KEGG Compound IDs, HMDB IDs, or ChEBI IDs. The mapping process is done by using AnnotationDbi [40] package internally.

Given a dataset with three types of omics (transcriptomics, proteomics, and metabolomics), multiGSEA maps the features to genes and then analyzes each data type independently using GSEA. After the analysis, each pathway has up to three p-values – one per data type. For each pathway, multiGSEA combines the three p-values using one of the four methods: Z-method or Stouffer's method or Fisher's combined probability test or Edgington's method [41]. The output of ~~the~~ the method includes four lists of p-values. Each pathway has three individual p-values (one per data type) and one aggregated p-value. The method also adjusts the p-values using False Discovery Rate.

10 MAPE

MAPE (Meta-Analysis for Pathway Enrichment analysis) [42] has an R package on CRAN. The input of this method includes multiple mRNA datasets and a set of gene lists (user-provided pathways). MAPE performs meta-analysis by combining the p-values obtained from each dataset. The combination can be performed at the gene-level (MAPE-G) or at the pathway level (MAPE-P). The two methods return two combined p-values – one at the gene-level and one at the pathway-level. Finally, MAPE-I combines these two p-values using the MinP method [43].

Briefly, MAPE-G first computes the t-statistic for each gene in each study and then calculate the p-value using permutation test. For each gene, MAPE-G then combines the individual p-values across multiple studies using the MaxP algorithm [44] to obtain on single p-value for the gene. Finally, given the combined p-values and statistics obtained from each gene, MAPE-G then performs a Kolmogorov-Smirnov (KS) test to calculate the p-value for each pathway. The KS-test used in this article is similar to that used in GSEA [45].

In contrast to MAPE-G, MAPE-P combines the p-values at the pathway-level. In each study, MAPE-P first computes the t-statistic for each gene and calculates the p-value for the gene using the permutation test. It then calculates the p-values for the pathways using the KS-test. After the p-values are calculated for each pathway in each study, MAPE-P then combines the p-values of a pathway across all studies using MaxP.

After the computation is done for both MAPE-G and MAPE-P, we obtain two p-values for each pathway. MAPE-I combines the two p-values using the MinP method to obtain one single p-value per pathway. The method outputs three lists of p-values – one for each of the three methods (MAPE-G, MAPE-P, and MAPE-I). The software also adjusts the p-values using the False Discovery Rate (FDR).

11 iPEAP

iPEAP (integrative Pathway Enrichment Analysis Platform) [46] is a Java software designed for pathway analysis using multi-omics data. At the time of writing this article, the software package is available for download from Google Drive. Supported data types include transcriptomics, proteomics, metabolomics, and SNP data. For each data type, the software utilizes one of the following methods to calculate the p-values of KEGG pathways: hypergeometric test (i.e., ORA), GSEA [45], Globaltest [47], SNP-GSEA [48], and Gowinda [49]. After performing pathway enrichment analysis on each data type separately using the selected algorithm, an aggregation algorithm is applied for integrating pathways enrichment scores and creating a unified ranked table of enriched pathways. Two rank aggregation methods are provided, RankAggreg [50] and RobustRankAggre [51]. Three measurements, Normalized discounted cumulative gain (NDCG) [52], Expected reciprocal rank (ERR) [53], and Proportion (P) [54] are considered for evaluating the ranking results. The NDCG considers the position of a relevant pathway in the aggregated ranked list. The ERR takes the user satisfaction from ranking results into consideration to evaluate the aggregated ranks. The last measurement, P computes the proportion of relevant pathways in the k first ranked pathways. Simple aggregation methods such as min, median and mean are also implemented.

12 IMPaLA

IMPaLA (Integrated Molecular Pathway-Level Analysis) [55] is a web-based tool that allows users to integrate gene expression with metabolomics data for pathway analysis. Genes and metabolites are linked through biochemical reactions and contained in many pathways. Accordingly, IMPaLA analyzes both types of data simultaneously. It combines the analysis of gene/protein and metabolite data using a comprehensive basis of existing biochemical pathways taken from 11 publicly available resources.

The website allows users to upload genes and/or metabolites in several identifiers namespaces. It will provide users two options of pathway analysis methods with the available pathways. The two methods are over-representation (ORA) or Wilcoxon enrichment analysis (WEA) [56]. Each of them also has different requirements for users to upload their datasets.

For ORA, the input data should be lists of significantly differentially expressed genes/proteins/metabolites, corresponding with the effect. These lists are resulted from differential expression analysis performed by the users. Users should upload the background lists of identifiers representing all measured genes/proteins and/or metabolites from the experiments. If no background lists are provided, all entities from existing pathways and annotated in the user-specified identifier namespaces are used as default background lists.

For WEA, input data are lists of all measured genes/proteins and/or metabolites without any differential expression analysis performed on them. Each entity contains one or a pair of numerical values. The values can be fold changes or average expression values for two different experiment conditions (e.g. normal and disease).

The pathway enrichment analysis is then performed on each individual datasets. The adjusted p-values (Q-values) are calculated using FDR. If there were no transcripts or no metabolites measured for a specific pathway, pathway's

p-value for that data type will be set to be 1. The pathway enrichment p-values from individual analysis are then combined by calculating their product value.

The output of IMPaLA is a table showing the pathways that contain at least one gene and/or metabolite from the input lists. Additionally, information about pathway name, source, size and overlap with the input entities is provided along with P-values calculated with appropriate statistical test for each pathway. Results can be sorted on any column by clicking on the appropriate column header, and can be downloaded as a tab-delimited file. By clicking on a pathway name, the user is guided to a summary web page at the original source database, which in most cases also shows a detailed pathway diagram.

13 GeneTrail2

GeneTrail2 [57] is available as a web-service. This method performs integrative pathway analysis using the following data types: transcriptomics, proteomics, miRNA, and genomics (SNP). The method allows users to analyze each input omics dataset independently and compare the results obtained from all analyses the end. Supported pathway databases include KEGG, Reactome, GO [58, 59], WikiPathways [60], BioCarta, NCI, PharmGKB, SMPDB and SignaLink [61]. Supported identifier mappings include Entrez ID, Uniprot and Ensemble. The input is either: (1) a list of differentially expressed (DE) genes, (2) a list of genes with computed scores (e.g., fold change or effect size), or (3) an expression matrix (e.g., microarray data) with groupings (e.g., diseases versus controls).

When users provide an expression matrix, GeneTrail2 allows them to perform differentially expression analysis at the gene level using one of the following methods: simple fold change, Z-score, signal-to-noise ratio, Pearson correlation, Spearman correlation, F-test, Welch's t-test, shrinkage t-test, dependent t-test, Wilcoxon rank-sum test, and Wilcoxon matched-pairs signed-ranks test. After the statistics of the genes are calculated, the method allows users to perform pathway analysis using one of the 10 following methods: weighted KS-test, unweighted KS test, Wilcoxon test, over-representation analysis (ORA), sum, mean, median, and maxmean statistics, as well as one and two sample t-test. For methods that require permutation, users can either choose to permute the genes or the sample labels. The tool also adjusts the p-values using one of the following methods: Benjamini-Hochberg FDR, Benjamini-Yekutieli FDR, Bonferroni FWER, Sidak FWER, Holm FWER, Holm-Sidak FWER, Finner FWER, Hochberg FWER.

For each dataset, the output is a list of p-values for the pathways. The significant pathways can be viewed online or exported as excel files (or compressed). For data integration, two view modes are available. The union mode displays pathways that are significant in at least one dataset. The intersection mode only displays pathways that are significant in all datasets. Using these modes, the user can effectively balance the sensitivity and specificity of the analysis.

14 mitch

mitch [62] is available as a Bioconductor R package. Overall, mitch is a framework for pathway analysis and data integration (multi-cohort and multi-omic integration). The input of mitch includes (1) pathway information uploaded in GMT format, (2) differentially expressed genes/features for multiple datasets/data types. Users need to upload a list of differential expressed features, together with their p-values and log (base 2) of fold changes. Given multiple datasets, mitch only keeps genes/features common across all datasets. By default, for each dataset, a differential score (D) is calculated for each feature using the following formula: $D = -log_{10}(p - value) \times sgn(log_2 FC)$. Users can also upload their own scores if they wish not to use the default scoring formula. Based on the activity scores, mitch ranks the features in each dataset. The rankings in all datasets are then merged into a single table, where rows are genes/features and columns are the datasets.

For each pathway, mitch divides the genes in the rank list into two groups: (1) genes that belong to the pathway and (2) the remaining genes. For each dataset, this will result in two vectors: ranking of genes belonging to the pathways and rankings of the remaining genes. Multivariate ANalysis Of VAriance (MANOVA) [63] is then applied to test for the difference between the two groups of genes in all datasets. If only one dataset is provided, then MANOVA is simply an ANOVA test. If the difference between the two groups is significant, this pathway is considered to be significantly enriched. Moreover, the enrichment score for each pathway is calculated in each dataset separately by the formula $s = 2(R_1 - R_2)/n$, where R_1 is the mean rank of genes in the pathway, R_2 is the mean rank of genes

not in the pathway, and n is the number of genes overall. The effect size S for each pathway is then calculated as the square root of the sum of squared s values across all datasets.

The output of mitch is a ranked list of pathways, together with their MANOVA p-values and adjusted p-values (using FDR). The table also includes other columns representing effect size, enrichment scores in every dataset as well as standard deviation (SD) of these enrichment scores across datasets. The results can be visualized in several plots in high-resolution PDF. Outputs contain scatterplots of DE scores derived from the directional p-value method, filled contour plots of ranked profiles, histogram of gene set sizes, scatter plot of effect size measured by S distance, and statistical significance measured adjusted p-values and a pairs plot of s values for all gene sets. In addition, detailed plots are generated for a specified number of gene sets according to the prioritization approach selected. These include pairwise filled contour plots, pairwise scatter plots, and violin plots of enrichments in each contrast.

15 ReactomeGSA

ReactomeGSA [64] is available as an R Bioconductor package as well as a web application. This tool allows users to perform pathway analysis using multiple datasets from the same or different types of omics. Supported data include microarray measurements, RNA-Seq raw and normalized read counts, proteomics spectral counts, and intensity-based quantitative data. By default, the software supports the analysis using the Reactome database. The input dataset can be uploaded manually or imported from Expression Atlas and Single Cell Expression Atlas. This tool also supports both discrete quantitative and continuous data. For both types, ReactomGSA has implemented some modules with the purpose of processing the raw data in terms of normalizing and transforming.

Given an input matrix, ReactomeGSA first maps the gene identifiers to human UniProt identifiers. Next, ReactomeGSA applies one of the three methods for pathway analysis: Camera through the "limma" package [65], PADOG through the "PADOG" package, and the single-sample gene set enrichment analysis (ssGSEA) [66] through the "GSVA" package [67]. It is notable that the differential expression analysis is done implicitly at the pathway level by applying these methods. Therefore, all the mentioned methods work directly with expression values. The parameters for the pathway analysis (such as the kernel to use for the ssGSEA analysis) are automatically chosen based on the data. Finally, the pathway analysis results are converted to Reactome's internal data format so that, this method can visualize the results using the PathwayBrowser module. The results from different analyses can be seen and interactively explored side by side.

The R package also supports the analysis of single-cell (scRNA-seq) data. For this type of data, the mean expression of genes within a cluster through either Seurat's [68] "AverageExpression" function or scater's [69] "aggregateAccrossCells" function, depending on the input object, is calculated. Then, all of the proposed analysis methods are applicable to this type of data. This results in one pathway-level expression value per cell cluster.

16 iODA

iODA (integrative Omics Data Analysis) [70] is a java software that peforms integrative pathway analysis using multi-omics and multi-cohort integration. This tool supports the analysis of transcriptome profiles (mRNA or miRNA expression data) and protein-DNA interactions (ChIP-Seq data). Gene/miRNA expression data can be imported as a matrix with two biological conditions, e.g., disease and control. The tools applies six different methods for differential analysis at the gene level: Least Sum of Ordered Subset Squared (LSOSS) [71], Cancer Outlier Profile Analysis (COPA) [72], Maximum Ordered Subset T-statistics (MOST) [73], Outlier Robust T-statistics (ORT) [74], Outlier Sum (OS) [75], and the t-test. The DE genes detected by at least four methods are considered as putative outliers. iOAD also calculates a metric that represents the accuracy of each method. Based on this, users can select the suitable method for the analysis. As a result, iODA produces the list of DE genes for pathway analysis. For ChIP-seq data, iODA uses MACS [76] to detect significant peaks of the protein binding site. Then, it uses PeakAnalyzer [77] tool to assign the binding sites to target genes.

Next, iODA uses two alternative strategies for pathway analysis. In the first approach, it intersects the genes obtained from the input datasets. These are considered DE genes and will be used to calculate the p-values of the pathways using the hypergeometric test (ORA). In the second approach, iODA performs pathway analysis for each dataset, and then intersects the significant pathways to obtain the final list of significant pathways. The tool also adjusts the p-values using FDR.

17 PARADIGM

PARADIGM [78] with available source code on Github is designed for multi-omics data integration and pathway analysis. The input includes matrices of multiple data types: gene expression, copy number variation, and proteins levels. Pathways are collected from National Cancer Institute (NCI) Pathway Interaction Database (PID).

PARADIGM converts each pathway into a distinct probabilistic model, represented as a factor graph with both hidden and observed states. In details, a pathway is first modeled as a directed acylic graph where nodes are genes and edges are defined as either positive or negative influence on the downstream nodes. Each gene is then modeled as a pathway diagram using a set of four different biological entities for the gene, which describe the DNA copies, mRNA and protein levels, and activity of the protein, respectively. This allows the incorporation of all supported types of omics data. Next, the method constructs a list of factors that represent various types of interactions across genes including transcription factors to targets, subunits aggregating in a complex, post-translational modification and sets of genes in a family performing redundant functions. Finally, these factors are added to specify the factor graph for each pathway.

PARADIGM takes observed experimental data, then calculates scores for all nodes (observed states) and specifies the probability of the node being active (hidden states). For each node score, a positive or negative value denotes how likely it is for the node to be active or inactive, respectively. The score is calculated to maximize the probability of the observed values. For a single node, a p-value is associated with each score of each sample such that each node can be tagged as significantly active (1), significantly inactive (-1), or not-significant (0). Finally, the pathways can be ranked based on the average number of samples in which significant activity is detected per node.

18 microGraphite

microGraphite [79] is available as an R function under the AGPL-3 license. The tool allows users to integrate miRNA with mRNA data for pathway analysis. User can download the R source code from the authors' website. In order to implement this code, two addition R packages *graphite* and *Clipper* [80] from BioConductor are required. The input for this method is an expression matrix, which is a combination of miRNA and mRNA expression matrices. Its headers are sample names and its row names are in gene IDs and miRNA IDs. The sample annotations are also needed to group the samples into different experimental conditions.

First, applying *graphite*, the pathway annotations are converted into gene–gene networks. Based on the resulted network, microGraphite expands the gene-based pathway graphs with miRNA-gene interactions derived from (i) validated miRNA-gene interactions from mirTarBase, miRecords, and a manual bilbilographic research; and (ii) predicted miRNA–target interactions from KEGG, Reactome, Nature Pathway Interaction (NCI), and Biocarta databases.

Next, microGraphite performs pathway analysis using the algorithm implemented in *Clipper*. The method follows a two-step strategy. In step 1, it selects pathways with covariance matrices or means that are significantly different between experimental conditions. In Step 2, on the selected pathways, it identifies portions of the pathway mostly associated with the phenotype. Based on graph decomposition, the pathway network is decomposed into fully-connected components, termed as 'cliques'. Each clique is analyzed independently (according to the test on the means and/or concentration matrices) to obtain a p-value that represents the statistical significance of the clique. Next, the method builds a junction tree – having cliques as nodes and satisfying the running intersection property stated as: for any cliques C1 and C2 in the tree, every clique on the path connecting C1 and C2 contains C1 $\cap$ C2.

Using the junction tree structure (that resembles the signal propagation of the pathway), the method computes the scores of every path in the graph. A path is a list of adjacent significant cliques (allowing a maximum of one gap, where a gap is defined as a non-significant clique). The score of a path is a function of all the p-values of the cliques contributing to the path. To avoid redundant path information, the *Clipper* approach would be performed on each significant pathway recursively. Finally, all the top-scored paths are combined to generate a meta-pathway. In the last step, the meta–pathway paths are analyzed and ranked according to their involvement in the phenotype. A sample permutation approach (permuting samples across groups) is used to estimate the significance levels of the test on the mean and on the concentration matrices either for the whole pathway or for the cliques.

The output of the method includes: (i) a list of pathways composed of genes and miRNAs associated to the phenotype, and (ii) a list of scored circuits composed by nodes, both genes and miRNAs and interactions that are strictly involved in the biological problem studied.

19 MOSClip

MOSClip (Multi-Omics Survival Clip) [81] is a standalone R package that exploits the topology of pathway annotations and integrates multi-omics data to identify pathways or pathway modules associated with right-censored survival data. MOSClip supports four types of omic input: (1) Expression (X - numerical matrix, data should be normalized and log-transformed), (2) Methylation (M - numerical, β value matrix), (3) mutational (U), and (4) copy number variation (CNV - C) should be in binary matrices (presence/absence of mutation/CNV or GISTIC thresholded data for CNV). Patient matching across the different omics is required. Moreover, the patient clinical data should be submitted since they are used to extract the Overall Survival (OS) and Progression Free Survival [82]. Another input is pathway data/gene set. Using pathway annotations, a graph structure of pathway is constructed as model G = (P, E) where P represents nodes (genes), and E represents edges (genes interaction). From this graph, all possible connected components are found, which refer to as pathway modules. Thus, multi-omics survival tests are performed either on a pathway or module level.

For each pathway or pathway module, MOSClip will go through the following process: (1) data filtering to keep only the genomic features that belong to the pathway/module; (2) dimension reduction on filtered data; (3) concatenate the data across data types by patient matching, and (4) perform survival analysis using the multivariate Cox proportional hazard model [83]. For dimension reduction, MOSClip uses PCA and hierarchical clustering on X and M. For U and C (binary), it summarizes the binary matrix with a sample binary vector having 1 if at least one gene in the pathway/module is mutated, amplified, or deleted and 0 otherwise, or a numeric vector that counts the number of altered genes in the pathway. The output is one Cox p-value for each pathway, and the coefficient for each variable is provided in step (3). MOSClip also provides different approaches to identify the genes most associated with survival: the absolute value of gene loadings is used in PCA, then the Kruskall–Wallis test [84] is used to compare measurements across patient groups for cluster analyses, and the three genes with the highest number of events are reported for binary data.

Furthermore, MOSClip provides several graphical tools to browse, manage and help interpretation of results, such as heat maps of p-values and prioritized genes, the radial plot of pathway frequencies across multiple omics, network visualization, omics combination, and survival annotation (Kaplan–Mayer curves and log-rank tests are used to stratify patients according to the combination of pathway/module omic variables).

20 IMPRes-Pro

IMPRes-Pro (Integrative MultiOmics Pathway Resolution) [85] is available as a web-based application that is able to detect biological pathways associated with specific diseases by using the integration of transcriptomics and proteomics data with the annotations from gene-protein interaction databases. IMPres-Pro utilizes the information from three public databases to construct a background network: (1) protein-protein interaction (PPI) from STRING database [86], (2) gene-gene interaction from KEGG, and (3) transcript factor (TF)-target gene network derived from TRRUST database [87]. The TF - target gene regulations work as a joint to connect PPI and pathway network.

IMPRes-Pro allows users to integrate gene expression data with protein data in one analysis. Users need to upload the data matrices, in which rows are genes/proteins and columns are samples. Gene and protein IDs should be in the ID format used in KEGG and STRING PPI databases, respectively. IMPRes-Pro also provides an ID converter tool to users. Each sample must exist in both data matrices (gene and protein expression matrices). The sample grouping information is also required. The grouping information can be different experimental conditions or different experimental groups at different time points. Thus, a differential expression analysis algorithm is applied to the expression input. In the use case of the case-control expression data, a t-test with a Bonferroni correction [88] is used while in the use case of a time-series expression data, maSigPro - a regression-based approach [89] is executed. Users can also provide the seed nodes (mandatory) and the target nodes (optional). The seed nodes must be protein nodes. The software uses differentially expressed genes and proteins as target nodes if target nodes are not provided.

Given the input, the method implements the step-wise active pathway detection algorithm on the background network. Starting from one or more seed nodes, the method explores all potential paths that include as many target nodes as possible. These paths then form a pathway network. Next, IMPRes-Pro uses shortest path algorithm [90] with a customized penalty function to achieve the optimal pathway network. In details, for each node, the method calculates its static penalty and dynamic penalty scores. For static penalty, if the case-control data is submitted, the method calculates the static penalty using node penalty formula: $penalty(v) = \frac{1}{-log_2(p-value(v))}$, where $p - value(v)$

is the p-value of node v retrieved from differential expression analysis using t-test. If the time-series expression data is used, the static penalty is calculated using edge penalty formula $penalty(e) = 1 - |correlation(V_1, V_2)|$, where $correlation(V_1, V_2)$ is the Pearson's correlation of two vectors V_1 and V_2 – the vectors of the time-series expression data of node v_1 and v_2. The node's static score is then updated iteratively. For dynamic penalty, it is calculated dynamically in the dynamic programming process. The algorithm stops when each node achieves a minimum involvement score, where involvement score is the sum of its static penalty and its dynamic penalty.

The final active pathway network is detected by truncating and backtracking. If a leaf node (a node with no downstream nodes) is not a target node, it is regarded as redundant and will be truncated. When a node's downstream nodes are all truncated, but the gene itself is not a target node, it will also be truncated. IMPRes-Pro returns the final active pathway network in the form of a tree structure graph rooted at the seed proteins.

21 KaPPA-View

KaPPA-View (Kazusa Plant Metabolic Pathway Viewer) [91] is a web-based tool that allows users to visualize metabolic and transcriptome data for the analysis of plant metabolic pathway maps. The input is quantitative values for transcripts and/or metabolites submitted by the user as a .csv file. At the time of writing this review, this tool has provided a collection of 150 Arabidopsis, rice, Lotus japonicus, and tomato metabolic pathway maps. Using pathway maps information from the internal library, KaPPA-View is able to overlay input gene-to-gene and/or metabolite-to-metabolite relationships as curves on a metabolic pathway map. However, it is possible for users to upload their pathway maps, as well.

The KaPPA-View server inserts colored symbols (color pathway indicators for overall changes to transcripts and metabolites in individual pathways, subclasses, and categories) corresponding to a defined metabolic process on the maps and returns them to the user's browser. The changes of individual genes or metabolites in the form of different colors can be seen interactively through pop-up windows. The presentation of data in this manner is helpful in displaying quantitative data changes for individual transcripts and metabolites between different experimental conditions on the same set of metabolic pathway maps. Consequently, this facilitates the discussion of gene function. Furthermore, gene-to-gene and/or metabolite-to-metabolite relationships such as co-expression correlations of genes can be displayed on the maps. This is the distinctive feature of KaPPA-View and will help users, for example, to analyze the relationships between metabolic genes and transcription factors that control their expressions (functions of a transcription factor that regulates genes on a metabolic pathway).

22 3Omics

3Omics [92] is a web-based application that supports visualization of multi-omics data using correlation networks and pathway enrichment analysis of metabolomics data. Supported data types include transcriptomics, proteomics, and metabolomics. Five major modules are provided in this tool: correlation and coexpression analysis, GO-based enrichment analysis, phenotypic analysis, and pathway enrichment analysis. Supported identifiers include Entrez IDs, UniprotKB IDs, and PubChem CIDs. The website also has a converter that converts gene name to ID. Each input file should consist of the transcript, protein, or metabolite IDs and their corresponding measurements (e. g., concentration or intensity levels). After uploading input files, the server automatically computes correlation coefficients, coexpression values, and pathway enrichment scores.

In the correlation analysis, the Pearson correlation coefficients of expression values of each pair of input omics data are computed. As a result, a correlation network, in which nodes are representative of omics IDs and edges show the correlation between them, is produced that is available for users. This analysis is applicable to different types of omics which are mentioned in the previous paragraphs. In the case of having missing omics data, the missing omics type is recovered by text mining of the iHOP [93] results. The coexpression analysis module provides heatmaps as output in which rows are the expression of input molecules, and columns are the expression differences between experimental groups (treatment/control groups or time-series experiments). Each cell in the resulting image is colorized based on the input expression value. In the phenotype analysis, the downloaded OMIM [94] data which specifies the related genes in the human genome with specific phenotypes is using in order to identify the genes and genetic disorders (the phenotype analysis is used for human phenotypic mapping).

Using metabolites input data, the pathway enrichment analysis component can be applied in two normal and enriched modes: the normal mode displays user-provided metabolites via simple metabolite mapping to a pathway from the KEGG pathway database, and the enriched mode, in which by two user-provided datasets (a metabolite set and a significantly changed metabolite set) and applying a hypergeometric test, the enriched pathways from KEGG and HumanCyc databases would be identified. GO-based functional enrichment analysis is performed through the DAVID [95] knowledgebase Application Platform Interface (API). Three independent GOs are included: (i) biological processes, (ii) cellular components, and (iii) molecular functions. The input transcripts are used in 3Omics to calculate the p-value and FDR (False Discovery Rate) of each GO term using a modified Fisher's exact test in the DAVID API. The enriched GO terms associated with the given Entrez Gene IDs are reported in 3Omics. Users can export network images in SVG or SIF formats.

23 InCroMAP

InCroMAP [96] is available as a stand-alone Java software that can be used on almost every operating system. This method was developed for multi-omics enrichment and pathway visualization. This tool supports different types of omics data (genomics, transcriptomics, proteomics, epigenomics, and metabolomics), which may come from various platforms. It supports automatic mapping between different metabolite identifiers (e.g., InChIKeys, HMDB, common synonyms) and different proteomic and transcriptomics annotations. Users can also upload user-defined mappings. When analyzing data using this software, any entries without InChiKey ID (i.e., information on compound classes) will be automatically discarded. It is because InCroMAP uses InChiKey as an internal identifier in order to merge four different compound databases (HMDB, LIPIDMAPS, KEGG, and PubChem).

In the pre-processing step, users need to normalize the data and perform differential analysis externally to calculate the gene-level statistics (e.g., p-values, fold changes, or log ratios). Then, the gene list and computed statistics can be uploaded to the software using a tabular format (similar to CSV format) which consists of platform-dependent meta-data columns containing appropriate identifiers (e.g., Affymetrix, EntrezGene, HMDB IDs) and data columns corresponding to the fold-changes or p-values. Another input file required from users is the pathway information obtained from pathway databases (such as KEGG, Reactome, BioCarta) in BioPAX format.

For each data type, InCroMAP first applies a threshold to identify the DE genes within each dataset (using fold change or p-values). Then the intersection or union of identified DE genes are considered for pathway enrichment analysis using ORA. The software adjusts the computed p-values using FDR. The output is a list of pathways and their nominal and adjusted p-values. There are two different visualizations of enriched pathways. One is the metabolic overview function of InCroMAP that generates an interactive global map of cellular metabolism, in which each subordinate metabolic pathway is colored according to the significance of its enrichment. The other is the integrated pathway-based visualization of data from multiple omics platforms. The tool enables users to browse through and visualize different KEGG pathways interactively.

24 Pathview

Pathview [97] is available as a Bioconductor's R package and a web application. This tool is mainly designed to provide a comprehensive view of a variety of biological data on pathway graphs. The format of input files can be either tab-delimited text (txt) or comma-separated values (CSV). Two different following categories can be imported as input in matrices or vectors format. First, gene data cover any data that map to unique gene IDs, including genes, transcripts, genomic loci, proteins, enzymes, and attributes. Second, compound data cover any data that map to unique compound IDs, including compounds, metabolites, drugs, small molecules, and their attributes. The pathway IDs can be selected (derived from the KEGG database) manually or automatically by the Pathview for integrative analysis. The functional Mapper module implemented in this tool can map 12 types of gene or protein IDs and 21 types of compound or metabolite-related IDs to standard KEGG gene or compound IDs and also map between these external IDs. For other types of IDs (for instance, Affymetrix microarray probe set IDs) not included in the common ID lists, Pathview's auxiliary functions will map user data to pathways when users manually provide the ID mapping data.

Pathview works with both numeric (e.g., expression levels) and categorical (e.g., gene or compound ID) data. In the numeric case, the GAGE [98] method is applied to the data to calculate the p-values of the pathways. For

categorical data, a hypergeometric test is applied. When both types of data are present, analysis is done on each data type separately, and then the results are combined into global statistics/p-values through meta-analysis. This tool outputs the pathway graphs with user data mapped in two views. First, the native KEGG view, in which the KEGG pathway diagram is used and the input transcriptomes and metabolites are laid. This view is considered more interpretable as all detailed meta-data such as input, output, connections, and tissue are available. The second view is Graphviz [99] view that provides better control over node/edge attributes and a better view of graph topology. The pathway analysis statistics will be returned, and all analysis results will be included in a zipped folder for download.

25 PaintOmics 3

PaintOmics 3 [100] is available as a website, and it supports the analysis of the following data types: gene expression, metabolomics, region-based omics (i.e., ChIP-seq, DNase-seq, ATAC-seq, Methyl-seq), and regulatory-based omics (miRNAs, transcription factors, or other factors). The tool allows users to integrate multi-omics data and perform pathway analysis using KEGG pathways. The input can be multiple files of the same or different types of supported omics in a tab-delimited format that contains the log fold change (e.g., disease vs. control) for each feature. Users can also input the list of DE genes directly. The tool automatically recognizes Ensembl, PDB, NCBI RefSeq, and Entrez ID. Users can also upload user-provided mapping. For mapping metabolites data to KEGG compounds, PaintOmics 3 generates a list of ranked candidates based on how similar their names are. Users can review and eventually change metabolite assignments if necessary. The website requires users to upload additional files for ID mapping for region-based and regulatory-based. With region-based data, the information of the chromosome, start and end position, and a quantification value for the regions must be provided, along with a GTF file with the reference genome annotation. PaintOmics 3 maps each region to its proximal gene(s) with the RGmatch [101] tool, which takes into account the relative position of the region with respect to the specific areas of the gene (such as promoter region, first exon, intronic areas). With regulatory-based omics, a tabulated file containing the associations between miRNAs/TF and their respective genes should be included in the input.

For each data type, the website first identifies the DE genes (by applying a threshold on the log fold change or provided by users) and then performs enrichment analysis using ORA. Then, for each pathway, the p-values from multiple data types are combined using Stouffer's or the weighted Fisher's method. The tool also adjusts the p-values using FDR. Users can view the result in the form of a table, in which rows are KEGG pathways and columns are their corresponding p-values for each data type as well as combined p-values.

PainOmics 3 provides three modules for visualizing the above analysis result. First, a pie chart and hierarchical structure show that KEGG pathways are organized in around seven main classifications (Cellular Processes, Drug Development, Environmental Information Processing, Genetic Information Processing, Human Diseases, Metabolism, and Organismal Systems) and/or over 50 secondary classifications. The second module is a pathways interaction network, in which the nodes represent pathways and edges indicate shared features among them. Finally, the last module allows users to explore individual pathways, showing the interactive pathway diagram, and the global heatmap contains information complementary to this pathway.

26 GSOA

GSOA (Gene Set Omic Analysis) [102] is available as an R script and a python-bash version. The tool supports the integration of multi-omics data for pathway analysis. The input includes multiple matrices (one matrix per data type), grouping information, and pathway information in the format of a GMT file. Supported data types include: gene expression, copy number variation (CNV), SNP, and other omics data. In each matrix, rows represent genomic features whole columns represent samples.

For each pathway, the omics measurements are filtered to keep only genes belong to that pathway. After filtering, if users input multiple omics data files, GSOA merges the data into a single data frame that includes whichever genes map to a given pathway for each omic type, even though different omic technologies may profile different genes. However, GSOA only considers samples that appear in all data types. Next, a classification method is applied to predict the class of each sample. By default, GSOA uses radial basis function support vector machine kernel (RBF SVM Kernel) [103] as classification algorithm and k-fold validation method with k = 5. Users can modify

these default parameters and clustering method as per their interests. After classification, the method assesses the prediction accuracy by calculating the area under the receiver operating characteristic curve (AUC). Relatively high AUC score means that the class of sample is accurately predicted. For example, an AUC score of 0.5 means that the classification is similar to a random guess. GSOA also performs cross-validation for randomly selected pathways to remove any correlation between pathway sizes and AUC values. Under the null hypothesis that the correlation exists, for each pathway, a p-value is calcualted as the fraction of AUC values that exceed the observed AUC value. The output of GSOA is a list of p-values ranked according to their nominal and FDR-adjusted p-values.

27 PathwayPCA

The PathwayPCA [104] is available as an R Bioconductor package for integrative pathway analysis using multiple types of molecular data (SNV, CNV, Proteomeics, miRNA, mRNA, Methylation data, Somatic Mutation data). The users need to provide three input files: (1) pathways collection in .gmt format, (2) expression data files (comma-separated .csv, fixed-width .fwf, or tab-delimited .txt) in which rows represent samples and columns represent genes, and (3) phenotype file in .csv format.

First, using the CreateOmics function, users can transform the input (pathways, expression, and phenotype) into a S4 data object. After the S4 object is prepared, users can call the functions AESPCA_pVals and SuperPCA_pVals to execute the methods AES-PCA [105] and SuperPCA [106, 107], respectively. The AES-PCA method is based on Adaptive, Elastic-net, Sparse PCA. This function first computes the latent variables for each data type and each pathway. Given a data type, the latent variables are tested against the phenotype, either using a regression model $g(phenotype) = \alpha + \beta PC1$ (default) where α and β are optimizable coefficients, or a link function g() that varies according to the response variable (i.e., Cox Proportional Hazards, identity, and logit link functions for survival, continuous, and binary response variables, respectively). Also, the number of permutations can be specified via *numReps* input argument. This function returns a named vector of permutation p-values. Finally, the obtained p-values are adjusted (using predefined methods including: Benjamini-Hochbergand FDR, Benjamini-Yekutieli FDR, adaptive Benjamini-Hochberg FDR, two-stage Benjamini-Hochberg FDR, Bonferroni FWER, Holm FWER, Hochberg FWER, Sidak single-step FWER and Sidak step-down FWER), and then sorted in a data frame.

PathwayPCA has extended the AES-PCA algorithm for integrative analysis of two omics data types (for example gene expression and proteomics data). In the proposed global test, after computing the first PCs for each type of omics separately for a pathway, the regression model $g(phenotype) = \alpha + \beta_1 PC1_{protein} + \beta_2 PC1_{RNA}$ is fitted (the g() function as mentioned above can differ based on the response variable), and a global test of the null hypothesis $H_0 : \beta = (\beta_1 \beta_2)^t = 0$ would be performed. Then, the joint effect of proteins and gene expressions in the pathway can be tested using a two degrees of freedom likelihood ratio test. However, we note that the package does not implement this integration. After identifying the significant pathways in each input omics dataset, by performing an inner join the significant pathways in all omics datasets are listed.

The SuperPCA is a supervised approach. In this method, first each input dataset are tested (using a cross validation approach) to identify the most relevant genes with the phenotype. Therefore, genes are filtered by applying a threshold. Next, silmilar to AES-PCA, the PCs for each pathway and each set of filtered genes are extracted. Lastly, the p-value for each pathway is computed using a two-component mixture of Gumbel extreme value distribution.

As a result of applying AESPCA_pVals and SuperPCA_pVals functions, a table including the analyzed pathways sorted by p-values with additional fields including pathway name, description, number of included features, and estimated False Discovery Rate would be returned. Also, PathwayPCA implements a range of functions for different purposes, including, getPathPCLs and SubsetPathwayData. The first function, getPathPCLs, with two required input arguments (the resulted output from AESPCA_pVals or SuperPCA_pVals function and a pathway name), returns two data frames, 1) a data frame of the samples and values for each principal components and 2) a data frame of the loading values corresponding to each gene. The SubsetPathwayData function, which has two mandatory inputs (the S4 object and a pathway name), returns a data frame contained the subset of assay data, by the matching gene symbols or IDs in the specified pathway, and the response data.

28 CPA

CPA (Consensus Pathway Analysis) [108] is a web-based platform for multi-cohort analysis. The input data for this tool include (i) pathways definition and (ii) gene data. This tool supports KEGG and GO pathway databases. Also, users can import their selected pathway collection in GMT format. For omics data, users can either import a list of DE genes, a list of genes and the corresponding fold changes, or a gene expression matrix. CPA maps all input genes into Entrez IDs to be compatible with the supported pathway databases.

Eight pathway analysis methods are included in CPA, which enable users to perform each on their input data. These are GSEA, GSA, FGSEA, PADOG, Impact Analysis, ORA/WebGestalt, KS-test and Wilcox-test. Users can choose each strategy to analyze their input data per desire. In addition, users are able to perform meta-analysis on multiple datasets and combine the resulted p-values using one of the Fisher, Stouffer's, addCLT, or minP methods.

After completing each analysis, a pathway-pathway graph is generated. Each node in this graph represents a significant pathway. Nodes are sliced based on the number of input datasets and the number of analysis methods performed on each input dataset. If the pathway is significant in a specific dataset and an applied method, the slice is colored. Otherwise, the slice remains white. The node's size corresponds to the number of genes within that pathway, and the number of DE genes specifies the thickness of the node's border. The details statistics related to each node and slice can be explored by hovering over each node. Users can interactively remove or add any number of pathways to this graph representation. Also, by selecting each node, a table is shown which includes information regarding genes in the pathway, the p-value and fold change in each dataset, and each pathway analysis method. All the analysis results besides the generated images are exportable.

29 pathwayMultiomics

PathwayMultiomics [109] is available as an R package. The package allows users to integrate analysis results obtained from multiple readouts (e.g., multiple datasets or multiple types of omics data). Note that this tool does not support pathway analysis using single-omics data, so users have to perform analysis on each readout (single-omics dataset) themselves by using any existing pathway analysis methods. After calculating the p-values for each pathway for each readout, users can provide the p-values for pathwayMultiomics to perform data integration. The input for this tool is a matrix of p-values with rows representing pathways and columns representing the different analyses.

The procedure to integrate input p-values is as follows. Given multiple p-values for a pathway – one for each readout – pathwayMultiomics starts with calculating a MinMax statistic for this pathway. Specifically, having p-values from all input datasets for a pathway, pathwayMultiomics first generates all possible combinations of any two p-values. For each pair, it then selects the maximum p-value. Next, among these selected p-values, pathwayMultiomics picks the minimum value as the MinMax statistic. In this case, the MinMax statistic is also known as the second order statistic (second smallest p-value) of the input p-values for each pathway. The statistical significance of the chosen MinMax statistics for each pathway is then assesed by utilizing an application of Beta distribution. That is, assuming all input p-values for a specific pathway are independent and uniformly distributed, the rth order statistic $p_{(r)}$ will follow a Beta distribution $\beta(\alpha = r, \beta = G - r + 1)$, where G is the number of p-values. Note that $r = 2$ in the case of MinMax statistic. The MinMax statistic is considered statistically significant if its $p_{(r)}$ value is less than a predefined threshold. Finally, the output of pathwayMultiomics is a table representing each pathway's computed MinMax statistic, p-value, and "driver omics". The driver omics are basically the pair of analyses that have the smallest p-values.

30 rPAC

The method rPAC (route-based pathway analysis for cohorts of gene expression datasets) [110] is designed to find the potential routes of a pathway that are significantly associated with the disease of interest. The input for rPAC includes: (1) a set of pathways in KGML format and (2) multiple gene expression matrices with genes in rows and samples as columns. A route can be either signaling route or effector route. Signaling route typically starts from ligand/receptor and follows a path to the primary transcriptome factor (TF). Effector route captures the cellular process where the primary TF binds to the target genes to regulate their expression.

Given the pathway file from KEGG (in KGML format), rPAC first extracts the nodes and edge information in which nodes include ligand, receptor, gene, and transcription factor whereas edges describe the relationships between genes. Next, rPAC constructs the pathway graphs based on the extracted information and then performs a Breadth-First Search (BFS) to identify potential routes. Only routes involving transcription factors are considered for further analysis. Each node k in route i is assigned an expected value E_{ik} based on its effect on activating the involved transcription factor (activation or inhibition):

$$E_{ik} = \begin{cases} 1, & \text{if } k = 1 \\ E_{i,k-1} * e_{ik}, & \text{otherwise} \end{cases}$$

where e_{ik} is $+1$ if the edge is activation and -1 in the case of inhibition of transcription factor.

In the next step, each node k in route i and sample j is evaluated and assigned a value v_{ijk}. If a node corresponds to a single gene, then v_{ijk} of this node equals to log2 ratio of the gene. In the second form, if a node is a bundle node which is a combination of multiple nodes, the v_{ijk} is computed as follows:

$$v_{ijk} = \begin{cases} \frac{\sum_{m=1}^{M} v_{mjk}}{M} * U_{ijk}, & U_{ijk} \geq U_{min} \\ \frac{\sum_{n=1}^{N} v_{njk}}{N} * (1 - U_{ijk}), & U_{ijk} < U_{min} \end{cases}$$

where M and N show the number of up and down-regulated nodes in the bundle, v_{mjk} and v_{njk} are their corresponding values. Also, $U_{ijk} = \frac{M}{n_k}$, where n_k equals the total number of nodes in bundle k. Furthermore, a bundle node can explicitly imply AND or OR interactions between membered nodes. In OR form, up/down-regulating of some of the nodes are enough to activate the transcription factor. On the contrary, the AND form requires all the nodes involved in an AND interaction to up/down-regulated to activate the transcription factor. U_{min} is empirically decided as 0.5 for AND bundle and 0.2 for the OR bundle.

Finally, rPAC assigns a score S_{ij} for each route i with respect to each sample j as $S_{ij} = \frac{1}{n_i} * \sum_{k=1}^{n_i} C_{ijk} * E_{ik}$, where n_i is the number of nodes in route i, and C_{ijk} is the contribution value of node k in route i for sample j. The activity score for each route ranges between -1 to $+1$ based on the down or up-regulation of that route. As a result, a matrix of computed scores for a specific route for a cohort of samples is produced. The p-value for each route within the same pathway is computed by testing a two-tailed hypothesis using a null distribution generated from sampling with replacement.

For each route, rPAC also calculates two metrics named Proportion of Significance (PS) and Average Route Score (ARS). The PS_i value of route i is calcualted as $PS_i = \frac{1}{J} * \sum_{j=1}^{J} I_{ij}$ where J is the number of samples in given cohor and $I_{ij} = 1$ if the p-value of the route i in sample j is less than the threshold (default value 0.05); otherwise $I_{ij} = 0$. The PS_i value represents the rate that a route is altered within a cohort ($0 \leq PS_i \leq 1$). The ARS_i value is calculated as $ARS_i = \frac{1}{J} * \sum_{j=1}^{J} S_{ij}$. The ARS_i value represents the average of yielded activity scores of a route within all samples in a cohort ($-1 \leq ARS_i \leq 1$).

31 clusterProfiler 4.0

clusterProfiler 4.0 [111] is available as an R package on Bioconductor. The software allows users to perform integrative pathway analysis using transcriptome and epigenome data. The method requires the following input: (i) one or multiple lists of differentially expressed genes, or lists of genes and their statistics (p-values and log fold change), and (ii) pathway database. For epigenome data, the authors of clusterProfier also provides an implementation of their previously developed method, 'ChIPseeker' [112], to map genomic regions to genes. The second input (pathway information) can be provided in the ".gmt" (gene matrix transpose) format. Alternatively, users can use embedded databases that include KEGG, GO and WikiPathways. Given that the input includes multiple lists of genes, users can perform both multi-omics integration and meta-analysis.

Given a single list of genes (and their statistics) and a pathway database, users can perform pathway analysis using ORA or GSEA using the following functions: *enrichGO, enrichKEGG, enrichMKEGG, enrichWP, enricher,*

gseGO, gseKEGG, gseMKEGG, gseWP, GSEA. The obtained p-values for the pathways can be adjusted using one of the following methods: Holm's, Bofferoni's, Hochberg's, Hommel's, Bonferroni–Holm, or Benjamini- Hochberg's false discovery rate (FDR). Given multiple lists of genes, users can perform analysis for each list and then compare and contrast the results using the *compareCluster* function. This function returns a table of pathways with multiple adjusted p-values – one for each input list. Users can further visualize the results in a side-by-side graph for comparison analysis.

32 Subpathway-GM

Subpathway-GM [113] is currently available as a function of the iSubpathwayMiner R package [114]. The method can perform integrative pathway analysis using transcriptome and metabolome data. The input of the method includes a list of differentially expressed genes (or genes of interest), and a list of metabolites. The method also provides embedded pathway information curated from the KEGG database.

Given the input of differentially expressed genes and metabolites, the method calculates the enrichment score for each pathway as the following process. First, it maps the interesting genes (first input) and metabolites (second input) to the nodes in the pathway graph. Especially, metabolites are directly mapped to metabolite nodes, whereas genes are mapped to enzyme nodes. The genes and metabolites are then referred to as 'signature nodes' of the pathway graph. Second, the method finds all signature subgraphs of the underlying pathway graph. To this aim, it searches for the shortest path between any two signature nodes. If the shortest path length is smaller than a specified threshold, then the two signature nodes and other nodes (non-signature nodes) belonging to the path are added to the same node-set. The method next extracts the corresponding subgraph in the pathway graph according to each node-set. The subgraphs with a number of nodes smaller than another threshold are then filtered out. Finally, the method calculates the enrichment score for each subgraph using the hypergeometric test. The default background is all possible genes and metabolites of the underlying species. The p-value of the pathway is the smallest p-value of its subgraphs. The method repeats the above process for all metabolic pathways to obtain the p-values for all pathways. Finally, Subpathway-GM outputs the pathways, subgraphs, and their p-values.

References

[1] Julia Feichtinger, Ramsay J McFarlane, and Lee D Larcombe. Cancerma: a web-based tool for automatic meta-analysis of public cancer microarray data. *Database*, 2012:bas055, 2012.

[2] Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A Fonseca, Robert Petryszak, Irene Papatheodorou, , Ugis Sarkans, and Alvis Brazma. Arrayexpress update–from bulk to single-cell expression data. *Nucleic Acids Research*, 47(D1):D711–D715, 2019.

[3] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Rolf N. Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets – 10 years on. *Nucleic Acids Research*, 39(suppl 1):D1005–D1010, 2011.

[4] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Bilal El Houdaigui, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Tiago Grego, Cristina Guijarro-Clarke, Leanne Haggerty, Anmol Hemrom, Thibaut Hourlier, Osagie G Izuogu, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, José Carlos Marugán, Thomas Maurel, Aoife C McMahon, Shamika Mohanan, Benjamin Moore, Matthieu Muffato, Denye N Oheh, Dimitrios Paraschas, Anne Parker, Andrew Parton, Irina Prosovetskaia, Manoj P Sakthivel, Ahamed I Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Emily Steed, Michal Szpak, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Brandon Walts, Andrea Winterbottom, Marc Chakiachvili, Ameya Chaubal, Nishadi De Silva, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth R IIsley, Nick Langridge, Jane E Loveland, Fergal J Martin, Jonathan M Mudge, Joanella Morales, Emily Perry, Magali Ruffier, John Tate,

David Thybert, Stephen J Trevanion, Fiona Cunningham, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891, 2021.

[5] Susan Tweedie, Bryony Braschi, Kristian Gray, Tamsin EM Jones, Ruth L Seal, Bethan Yates, and Elspeth A Bruford. Genenames. org: the hgnc and vgnc resources in 2021. *Nucleic Acids Research*, 49(D1):D939–D946, 2021.

[6] GO. Gene Ontology. Technical report, Gene Ontology Consortium, 2001. http://www.geneontology.org/.

[7] Gordon K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[8] Samuel A. Stouffer, Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams Jr. *The American Soldier: Adjustment during army life*, volume 1. Princeton University Press, Princeton, 1949.

[9] Jianguo Xia, Christopher D Fjell, Matthew L Mayer, Olga M Pena, David S Wishart, and Robert EW Hancock. INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Research*, 41(W1):W63–W70, 2013.

[10] Marta Paczkowska, Jonathan Barenboim, Nardnisa Sintupisut, Natalie S. Fox, Helen Zhu, Diala Abd-Rabbo, Miles W. Mee, Paul C. Boutros, PCAWG Drivers and Functional Interpretation Working Group, Jüri Reimand, and PCAWG Consortium. Integrative pathway enrichment analysis of multivariate omics data. *Nature Communications*, 11:735, 2020.

[11] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS ONE*, 5(11):e13984, 2010.

[12] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, Kristina Hanspers, Ruth Isserlin, Ryan Kelley, Sarah Killcoyne, Samad Lotia, Steven Maere, John Morris, Keiichiro Ono, Vuk Pavlovic, Alexander R. Pico, Aditya Vailaya, Peng-Liang Wang, Annette Adler, Bruce R. Conklin, Leroy Hood, Martin Kuiper, Chris Sander, Ilya Schmulevich, Benno Schwikowski, Guy J. Warner, Trey Ideker, and Gary D. Bader. Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, 2(10):2366–2382, 2007.

[13] Alexander Kaever, Manuel Landesfeind, Kirstin Feussner, Alina Mosblech, Ingo Heilmann, Burkhard Morgenstern, Ivo Feussner, and Peter Meinicke. MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics*, 11(3):764–777, 2015.

[14] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society B*, 57(1):289–300, 1995.

[15] Peter D. Karp, Richard Billington, Ron Caspi, Carol A. Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M. Keseler, Markus Krummenacker, Peter E. Midford, Quang Ong, Wai Kit Ong, Suzanne M. Paley, and Pallavi Subhraveti. The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, 20(4):1085–1093, 2019.

[16] Rainer Breitling, Anna Amtmann, and Pawel Herzyk. Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, 5:34, 2004.

[17] Frank J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[18] Ronald A. Fisher. *Statistical methods for research workers.* Oliver & Boyd, Edinburgh, 1925.

[19] Tin Nguyen, Rebecca Tagett, Michele Donato, Cristina Mitrea, and Sorin Draghici. A novel bi-level meta-analysis approach Applied to biological pathway analysis. *Bioinformatics*, 32(3):409–416, 2016.

[20] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

[21] Adi L Tarca, Sorin Drăghici, Gaurav Bhatti, and Roberto Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13:136, 2012.

[22] Sorin Draghici, Purvesh Khatri, Adi L. Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.

[23] Tin Nguyen, Cristina Mitrea, Rebecca Tagett, and Sorin Draghici. DANUBE: Data-Driven Meta-ANalysis Using UnBiased Empirical Distributions—Applied to Biological Pathway Analysis. *Proceedings of the IEEE*, 105(3):496–515, 2016.

[24] Le Shu, Yuqi Zhao, Zeyneb Kurt, Sean Geoffrey Byars, Taru Tukiainen, Johannes Kettunen, Luz D Orozco, Matteo Pellegrini, Aldons J Lusis, Samuli Ripatti, Bin Zhang, Michael Inouye, Ville-Petteri Mäkinen, and Xia Yang. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics*, 17:874, 2016.

[25] Diana Diaz, Michele Donato, Tin Nguyen, and Sorin Draghici. MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping. In *The Pacific Symposium on Biocomputing 2017*, pages 390–401. World Scientific, 2017.

[26] Sheng-Da Hsu, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou, Chao-Fang Chu, Hsi-Yuan Huang, Ching-Min Lin, Shu-Yi Ho, Ting-Yan Jian, Feng-Mao Lin, Tzu-Hao Chang, Shun-Long Weng, Kuang-Wen Liao, I.-En Liao, Chun-Chi Liu, and Hsien-Da Huang. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, 42(D1):D78–D85, January 2014.

[27] Boya Xie, Qin Ding, and Di Wu. Targetscan 6.2 data download. `http://www.targetscan.org`, 2012.

[28] Sebastian Canzler and Jörg Hackermüller. multiGSEA: A GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinformatics*, 21:561, 2020.

[29] Gabriele Sales, Enrica Calura, Duccio Cavalieri, and Chiara Romualdi. g raphite-a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):1–12, 2012.

[30] M Whirl-Carrillo, EM McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, RB Altman, and Teri E Klein. Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012.

[31] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic Acids Research*, 37(suppl_1):D674–D679, 2009.

[32] Pedro Romero, Jonathan Wagg, Michelle L Green, Dale Kaiser, Markus Krummenacker, and Peter D Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6:R2, 2005.

[33] Timothy Jewison, Yilu Su, Fatemeh Miri Disfany, Yongjie Liang, Craig Knox, Adam Maciejewski, Jenna Poelzer, Jessica Huynh, You Zhou, David Arndt, Yannick Djoumbou, Yifeng Liu, Lu Deng, An Chi Guo, Beomsoo Han, Allison Pon, Michael Wilson, Shahrzad Rafatnia, Philip Liu, and David S. Wishart. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Research*, 42(D1):D478–D484, 2014.

[34] Huaiyu Mi, Anushya Muruganujan, and Paul D Thomas. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(D1):D377–D386, 2012.

[35] Darryl Nishimura. Biocarta. *Biotech Software and Internet Report*, 2(3):117–120, 2001.

[36] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[37] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2021.

[38] Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951, 2019.

[39] Lisa Matthews, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(suppl_1):D619–D622, 2009.

[40] Hervé Pagès, M Carlson, S Falcon, and N Li. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. *R package version 1.0*, 2019.

[41] Eugene S. Edgington. An additive method for combining probability values from independent experiments. *The Journal of Psychology*, 80(2):351–363, 1972.

[42] Kui Shen and George C Tseng. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323, 2010.

[43] Leonard H. C. Tippett. *The methods of statistics*. Williams & Norgate, London, 1931.

[44] Bryan Wilkinson. A statistical consideration in psychological research. *Psychological Bulletin*, 48(2):156, 1951.

[45] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding of The National Academy of Sciences*, 102(43):15545–15550, 2005.

[46] Haoqi Sun, Haiping Wang, Ruixin Zhu, Kailin Tang, Qin Gong, Juan Cui, Zhiwei Cao, and Qi Liu. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics*, 30(5):737–739, 2014.

[47] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[48] Kai Wang, Mingyao Li, and Maja Bucan. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283, 2007.

[49] Robert Kofler and Christian Schlötterer. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, 28(15):2084–2085, 2012.

[50] Vasyl Pihur, Susmita Datta, and Somnath Datta. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics*, 23(13):1607–1615, 2007.

[51] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

[52] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[53] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630, 2009.

[54] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. Frank: a ranking method with fidelity loss. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 383–390, 2007.

[55] Atanas Kamburov, Rachel Cavill, Timothy MD Ebbels, Ralf Herwig, and Hector C Keun. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, 27(20):2917–2918, 2011.

[56] James Adjaye, John Huntriss, Ralf Herwig, Alia BenKahla, Thore C Brink, Christoph Wierling, Claus Hultschig, Detlef Groth, Marie-Laure Yaspo, Helen M Picton, Roger G Gosden, and Hans Lehrach. Primary Differentiation in the Human Blastocyst: Comparative Molecular Portraits of Inner Cell Mass and Trophectoderm Cells. *Stem Cells*, 23(10):1514–1525, 2005.

[57] Daniel Stöckel, Tim Kehl, Patrick Trampert, Lara Schneider, Christina Backes, Nicole Ludwig, Andreas Gerasch, Michael Kaufmann, Manfred Gessler, Norbert Graf, Eckart Meese, Andreas Keller, and Hans-Peter Lenhof. Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, 32(10):1502–1508, 01 2016.

[58] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[59] The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.

[60] Thomas Kelder, Martijn P van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307, 2012.

[61] Dávid Fazekas, Mihály Koltai, Dénes Türei, Dezső Módos, Máté Pálfy, Zoltán Dúl, Lilian Zsákai, Máté Szalay-Bekő, Katalin Lenti, Illés J Farkas, Tibor Vellai, Péter Csermely, and Tamás Korcsmáros. Signalink 2–a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology*, 7(1):1–15, 2013.

[62] Antony Kaspi and Mark Ziemann. mitch: multi-contrast pathway enrichment for multi-omics and single-cell profiling data. *BMC Genomics*, 21:447, 2020.

[63] Chen-An Tsai and James J Chen. Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, 25(7):897–903, 2009.

[64] Johannes Griss, Guilherme Viteri, Konstantinos Sidiropoulos, Vy Nguyen, Antonio Fabregat, and Henning Hermjakob. ReactomeGSA-Efficient Multi-Omics Comparative Pathway Analysis. *Molecular & Cellular Proteomics*, 19(12):2115–2125, 2020.

[65] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 01 2015.

[66] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Edmond M. Chan, Martin L. Sos, Kathrin Michel, Craig Mermel, Serena J. Silver, Barbara A. Weir, Jan H. Reiling, Qing Sheng, Piyush B. Gupta, Raymond C. Wadlow, Hanh Le, Sebastian Hoersch, Ben S. Wittner, Sridhar Ramaswamy, David M. Livingston, David M. Sabatini, Matthew Meyerson, Roman K. Thomas, Eric S. Lander, Jill P. Mesirov, David E. Root, D. Gary Gilliland, Tyler Jacks, and William C. Hahn. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112, 2009.

[67] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14:7, 2013.

[68] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[69] Davis J McCarthy, Kieran R Campbell, Aaron TL Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 2017.

[70] Chunjiang Yu, Xin Qi, Yuxin Lin, Yin Li, and Bairong Shen. iODA: An integrated tool for analysis of cancer pathway consistency from heterogeneous multi-omics data. *Journal of Biomedical Informatics*, 112:103605, 2020.

[71] Yupeng Wang and Romdhane Rekaya. LSOSS: Detection of Cancer Outlier Differential Gene Expression. *Biomarker Insights*, 5:BMI–S5175, 2010.

[72] James W MacDonald and Debashis Ghosh. COPA—cancer outlier profile analysis. *Bioinformatics*, 22(23):2950–2951, 2006.

[73] Heng Lian. MOST: detecting cancer differential gene expression. *Biostatistics*, 9(3):411–418, 2008.

[74] Baolin Wu. Cancer outlier differential gene expression detection. *Biostatistics*, 8(3):566–575, 2007.

[75] Robert Tibshirani and Trevor Hastie. Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1):2–8, 2007.

[76] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9:R137, 2008.

[77] Mali Salmon-Divon, Heidi Dvinge, Kairi Tammoja, and Paul Bertone. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics*, 11:415, 2010.

[78] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–i245, 2010.

[79] Enrica Calura, Paolo Martini, Gabriele Sales, Luca Beltrame, Giovanna Chiorino, Maurizio D'Incalci, Sergio Marchini, and Chiara Romualdi. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Research*, 42(11):e96, 2014.

[80] Paolo Martini, Gabriele Sales, M Sofia Massa, Monica Chiogna, and Chiara Romualdi. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research*, 41(1):e19–e19, 2013.

[81] Paolo Martini, Monica Chiogna, Enrica Calura, and Chiara Romualdi. MOSClip: multi-omic and survival pathway analysis for the identification of survival associated gene and modules. *Nucleic Acids Research*, 47(14):e80–e80, 2019.

[82] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, Larsson Omberg, Denise M. Wolf, Craig D. Shriver, and Vesteinn Thorsson. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

[83] David R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[84] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[85] Yuexu Jiang, Duolin Wang, Dong Xu, and Trupti Joshi. IMPRes-Pro: A high dimensional multiomics integration method for in silico hypothesis generation. *Methods*, 173(1):16–23, 2020.

[86] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi .P Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian Von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43:D447–452, 2015.

[87] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, Sungho Lee, Byunghee Kang, Dabin Jeong, Yaeji Kim, Hyeon-Nae Jeon, Haein Jung, Sunhwee Nam, Michael Chung, Jong-Hoon Kim, and Insuk Lee. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380–D386, 2018.

[88] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[89] María José Nueda, Sonia Tarazona, and Ana Conesa. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, 30(18):2598–2602, 06 2014.

[90] E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1:269–271, 1959.

[91] Toshiaki Tokimatsu, Nozomu Sakurai, Hideyuki Suzuki, Hiroyuki Ohta, Kazuhiko Nishitani, Tanetoshi Koyama, Toshiaki Umezawa, Norihiko Misawa, Kazuki Saito, and Daisuke Shibata. KaPPA-View. A Web-Based Analysis Tool for Integration of Transcript and Metabolite Data on Plant Metabolic Pathway Maps. *Plant Physiology*, 138(3):1289–1300, 2005.

[92] Tien-Chueh Kuo, Tze-Feng Tian, and Yufeng Jane Tseng. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology*, 7:64, 2013.

[93] Jose M Fernandez, Robert Hoffmann, and Alfonso Valencia. ihop web services. *Nucleic Acids Research*, 35(suppl_2):W21–W26, 2007.

[94] Ada Hamosh, Alan F Scott, Joanna Amberger, David Valle, and Victor A McKusick. Online mendelian inheritance in man (omim). *Human mutation*, 15(1):57–61, 2000.

[95] Brad T Sherman and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44, 2009.

[96] Johannes Eichner, Lars Rosenbaum, Clemens Wrzodek, Hans-Ulrich Häring, Andreas Zell, and Rainer Lehmann. Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software. *Journal of Chromatography B*, 966:77–82, 2014.

[97] Weijun Luo, Gaurav Pant, Yeshvant K Bhavnasi, Steven G Blanchard Jr, and Cory Brouwer. Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Research*, 45(W1):W501–W508, 2017.

[98] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10:161, 2009.

[99] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. Graphviz—Open Source Graph Drawing Tools. In *International Symposium on Graph Drawing*, pages 483–484. Springer, 2001.

[100] Rafael Hernández-de Diego, Sonia Tarazona, Carlos Martínez-Mira, Leandro Balzano-Nogueira, Pedro Furió-Tarí, Georgios J Pappas Jr, and Ana Conesa. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Research*, 46(W1):W503–W509, 2018.

[101] Pedro Furió-Tarí, Ana Conesa, and Sonia Tarazona. Rgmatch: matching genomic regions to proximal genes in omics data integration. *BMC Bioinformatics*, 17(15):1–10, 2016.

[102] Shelley M MacNeil, William E Johnson, Dean Y Li, Stephen R Piccolo, and Andrea H Bild. Inferring pathway dysregulation in cancers from multiple types of omic data. *Genome Medicine*, 7:61, 2015.

[103] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks and Learning Systems*, 10(5):988–999, 1999.

[104] Gabriel J Odom, Yuguang Ban, Lizhong Liu, Xiaodian Sun, Alexander R Pico, Bing Zhang, Lily Wang, and Xi Chen. pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection. *bioRxiv*, page 615435, 2019.

[105] Xi Chen. Adaptive Elastic-Net Sparse Principal Component Analysis for Pathway Association Testing. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

[106] Xi Chen, Lily Wang, Bo Hu, Mingsheng Guo, John Barnard, and Xiaofeng Zhu. Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic epidemiology*, 34(7):716–724, 2010.

[107] Xi Chen, Lily Wang, Jonathan D Smith, and Bing Zhang. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, 24(21):2474–2481, 2008.

[108] Hung Nguyen, Duc Tran, Jonathan M. Galazka, Sylvain V. Costes, Afshin Beheshti, Sorin Draghici, and Tin Nguyen. CPA: A web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Research*, 49(W1):W114–W124, 2021.

[109] Gabriel J. Odom, Antonio Colaprico, Tiago Chedraoui Silva, Xi Steven Chen, and Lily Wang. PathwayMultiomics: An R Package for Efficient Integrative Analysis of Multi-Omics Datasets With Matched or Un-matched Samples. *Frontiers in Genetics*, 12:783713, 2021.

[110] Pujan Joshi, Brent Basso, Honglin Wang, Seung-Hyun Hong, Charles Giardina, and Dong-Guk Shin. rPAC: Route based pathway analysis for cohorts of gene expression data sets. *Methods*, 198:76–87, 2022.

[111] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, LI Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021.

[112] Guangchuang Yu, Li-Gen Wang, and Qing-Yu He. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383, 2015.

[113] Chunquan Li, Junwei Han, Qianlan Yao, Chendan Zou, Yanjun Xu, Chunlong Zhang, Desi Shang, Lingyun Zhou, Chaoxia Zou, Zeguo Sun, Jing Li, Yunpeng Zhang, Haixiu Yang, Xu Goa, and Xia Li. Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Research*, 41(9):e101, 2013.

[114] Chunquan Li, Xia Li, Yingbo Miao, Qianghu Wang, Wei Jiang, Chun Xu, Jing Li, Junwei Han, Fan Zhang, Binsheng Gong, and Liangde Xu. SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Research*, 37(19):e131–e131, 2009.